Report & Analysis

Introduction

This project aims to build a machine-learning model to predict soccer match outcomes using historical data. The trained model is then applied to simulate the 2018 FIFA World Cup multiple times, and the probability of different teams winning the tournament is determined.

Dataset and Feature Engineering

The dataset included 30,000+ international matches (1950-2017), and the additional feature of teams and World Cup qualifiers. Some key feature engineering steps are:

- Merging confederation data from teams.csv to matches.
- Handling missing values, assuming missing penalties as zero.
- Creating new features:
    - Goal difference (team1score - team2score)
    - Match result encoding: 1 (win), 0 (draw), -1 (loss)
    - Win percentages for both teams
    - Label encoding for teams

Model Training

A Random Forest Classifier was selected for the match outcome prediction because it can extract complex relations among the features and handle missing data well. The approach used was pipeline-based with:

- Preprocessing: One-hot encoding categorical features, standard scaling numeric features.
- Feature Selection: Team encodings, win percentages, goal difference, confederation, tournament year.
- Training: Data was split into 80% training, and 20% testing, which resulted in a high test set accuracy.

Tournament Simulation

Simulate the 2018 FIFA World Cup 1,000 times with the trained model to estimate the most wins. The simulation is based on a knockout-style tournament. It selects winners through:

- Model-predicted match results
- Aggregation of results for statistical analysis

Results & Insights

- Predictions were logical and aligned with historical trends.
- Most winning teams have consistently dominated the simulations.
- Biases in the model: The model lacks real-time player form or squad strength.

Improvements:

- Can add venue effects like home teams and away teams.
- Deep learning with richer feature extraction
- We can also improve accuracy by adding more relevant features like the current inform team.